CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD
COLLEGE

Admin

- No office hours Monday Oct 9
- Tuesday Oct 10 during lab we will continue review
- May have extra zoom office hours/meeting times on Wed

- **Midterm 1 handed out today**
  - Do not open until you are ready to take it!
  - 3 hour time limit
  - Due Thursday at the beginning of class (Oct 7)

Data Science and the Liberal Arts
A View from Down Under

Daniel Russo-Batterham
Senior Research Data Specialist
University of Melbourne

Thursday, October 5, 2023
4:15pm in Lutnick 232

TODAY!

Sponsored by the Music Department, TriCo Data Science, and HAVERFORD LIBRARIES

Are you hoping to break into Biotech & Pharma?

Hear from Tri-Co alumni in the biotech and pharma fields discuss career paths, required skills and how to break into these competitive industries! Alumni from the below organizations will participate in this interactive panel and networking event:

- **Eli Lilly**
- **IQVIA**
- **Pfizer**
- **Syncro Medical**

Event Structure:

TODAY!
(Lutnick 200)

- 6:30-7:00pm: Panel Q&A
- 7:00-8:00pm: Open networking with Tri-Co alumni!

🍪**INSOMNIA COOKIES WILL BE PROVIDED!**🍪

"Human-Centered Machine Intelligence" with Leqi Liu '17

Wednesday, October 11, 2023
4:30 p.m. to 5:30 p.m.
Lutnick Library
LUT 200 Instruction and Events
+ Add to Calendar

Why do we have a exam?

- Process of synthesizing the material on your own is essential
- Preparing the “study sheet” is designed to facilitate that process
- Review in class today and in lab on Tuesday (working through midterm practice questions)

Outline for Oct 5

- Review
  - Linear regression
  - Gradient descent
  - Matrix/vector form of Lab 3
  - Classification
  - Single feature models / decision trees
  - Evaluation metrics

Matrix / Vector form of SGD

while not converged:
#shuffle the data.
for $i = 1, 2, \ldots, n$:
sort indices
4, 7, 2, 1 ... instead

check cost not changing

$$\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix} \leftarrow \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix} - \alpha \left(h_{\vec{w}}(\vec{x}_i) - y_i \right) \begin{bmatrix} x_{i,0} \\ x_{i,1} \\ \vdots \\ x_{i,p} \end{bmatrix}$$

Scalar

$h_{\vec{w}}(\vec{x})$

$= \vec{w} \cdot \vec{x} = w_0 x_0 + w_1 x_1 \cdots w_p x_p$ (fake 1)

predict$(X, \vec{w})$

$$X\vec{w} = \begin{bmatrix} & & & \\ x_{i0} & x_{i1} & \cdots & x_{ip} \\ & & & \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_i \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

$n \times (p+1)$ $\quad (p+1) \times 1$ $\quad n \times 1$

cost$(X, \vec{y}, \vec{w})$

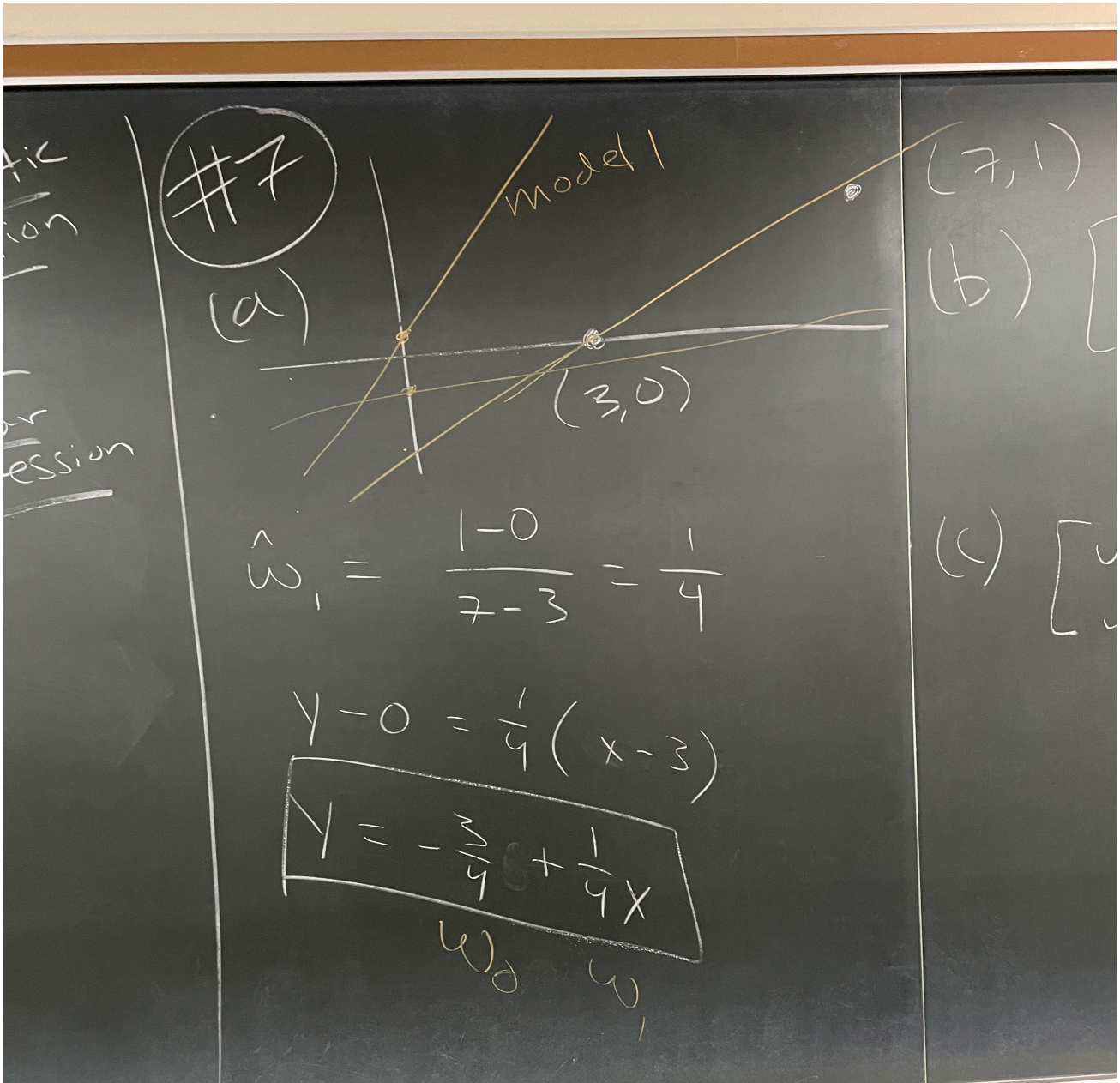
$$J(\vec{w}) = \frac{1}{2} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$$= \frac{1}{2} (\hat{\vec{y}} - \vec{y}) \cdot (\hat{\vec{y}} - \vec{y})$$

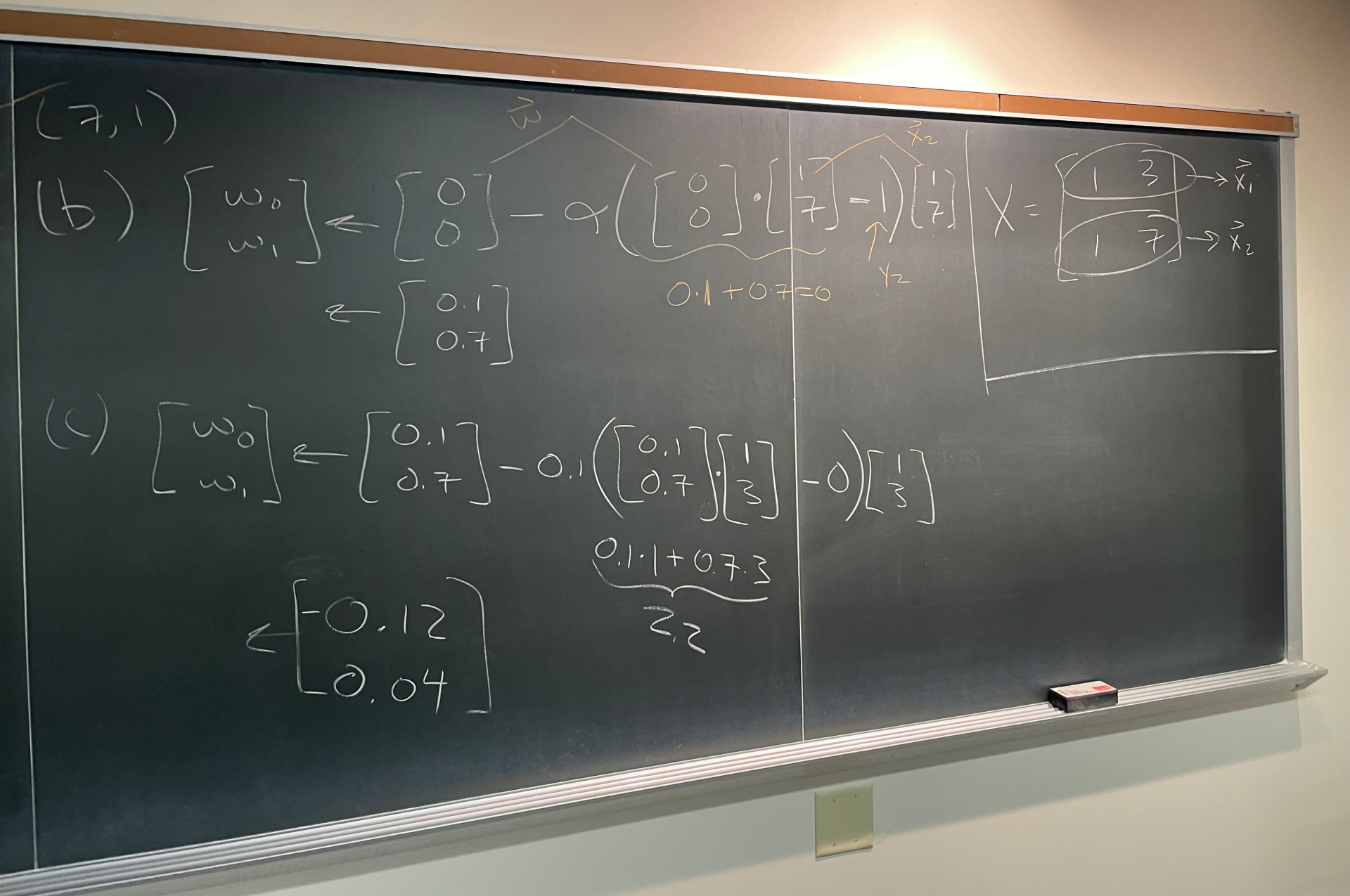
$$= \frac{1}{2} (X\vec{w} - \vec{y}) \cdot (X\vec{w} - \vec{y})$$

Page 3 #7

Handout 10, #7


#7
(a)
Model 1
(3,0)
$\hat{\omega}_1 = \frac{1-0}{7-3} = \frac{1}{4}$
$y - 0 = \frac{1}{4}(x-3)$
$y = -\frac{3}{4} + \frac{1}{4}x$
ω_0 ω_1
(7,1)
(b)
(c)

Handout 10, #7


(7, 1)
(b)
$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \leftarrow \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \alpha \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 7 \end{bmatrix} - 1 \right) \begin{bmatrix} 1 \\ 7 \end{bmatrix}$
$\leftarrow \begin{bmatrix} 0.1 \\ 0.7 \end{bmatrix}$
$0 \cdot 1 + 0 \cdot 7 = 0$
y_2
$\vec{x}_2$
$\vec{w}$
$X = \begin{bmatrix} 1 & 3 \\ 1 & 7 \end{bmatrix}$
$\vec{x}_1$
$\vec{x}_2$
(c)
$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \leftarrow \begin{bmatrix} 0.1 \\ 0.7 \end{bmatrix} - 0.1 \left(\begin{bmatrix} 0.1 \\ 0.7 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 3 \end{bmatrix} - 0 \right) \begin{bmatrix} 1 \\ 3 \end{bmatrix}$
$0.1 \cdot 1 + 0.7 \cdot 3$
2.2
$\leftarrow \begin{bmatrix} -0.12 \\ 0.04 \end{bmatrix}$

Time varying α | SGD

(α: learning rate) step size

$t = 1$

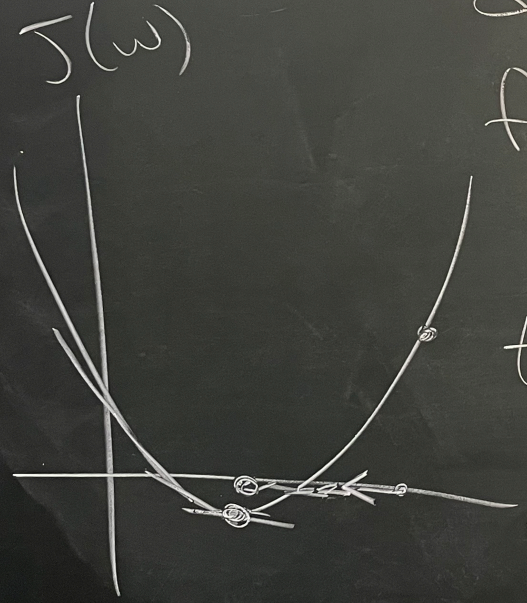
while not converged:

$\alpha = \frac{1}{t}$

for $i = 1 \dots n$

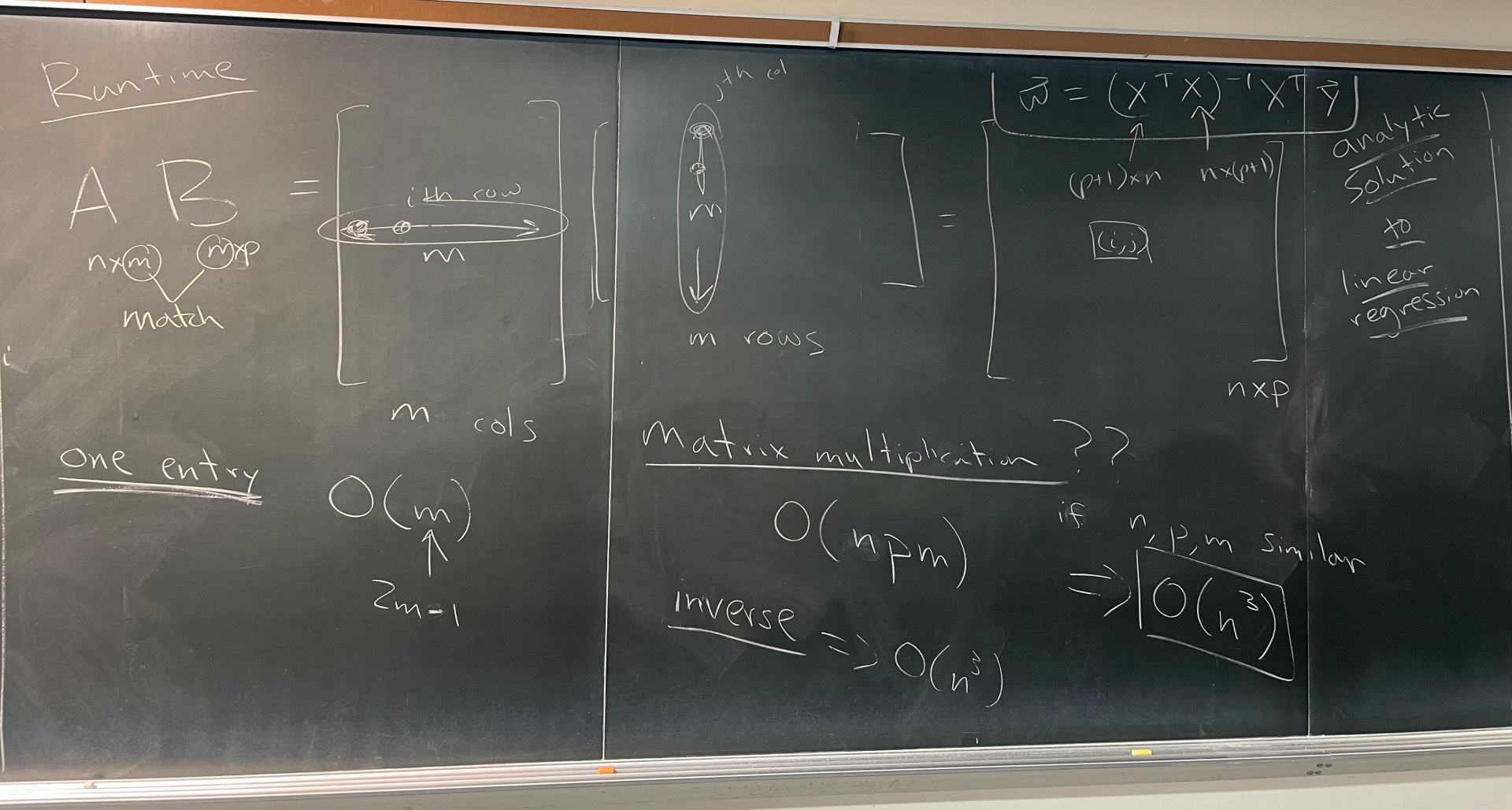
$\vec{w} \leftarrow \vec{w} - \alpha \underbrace{(h_{\vec{w}}(x_i) - y_i)}_{\text{derivative of wrt } \vec{w}} \vec{x}_i$

$t += 1$

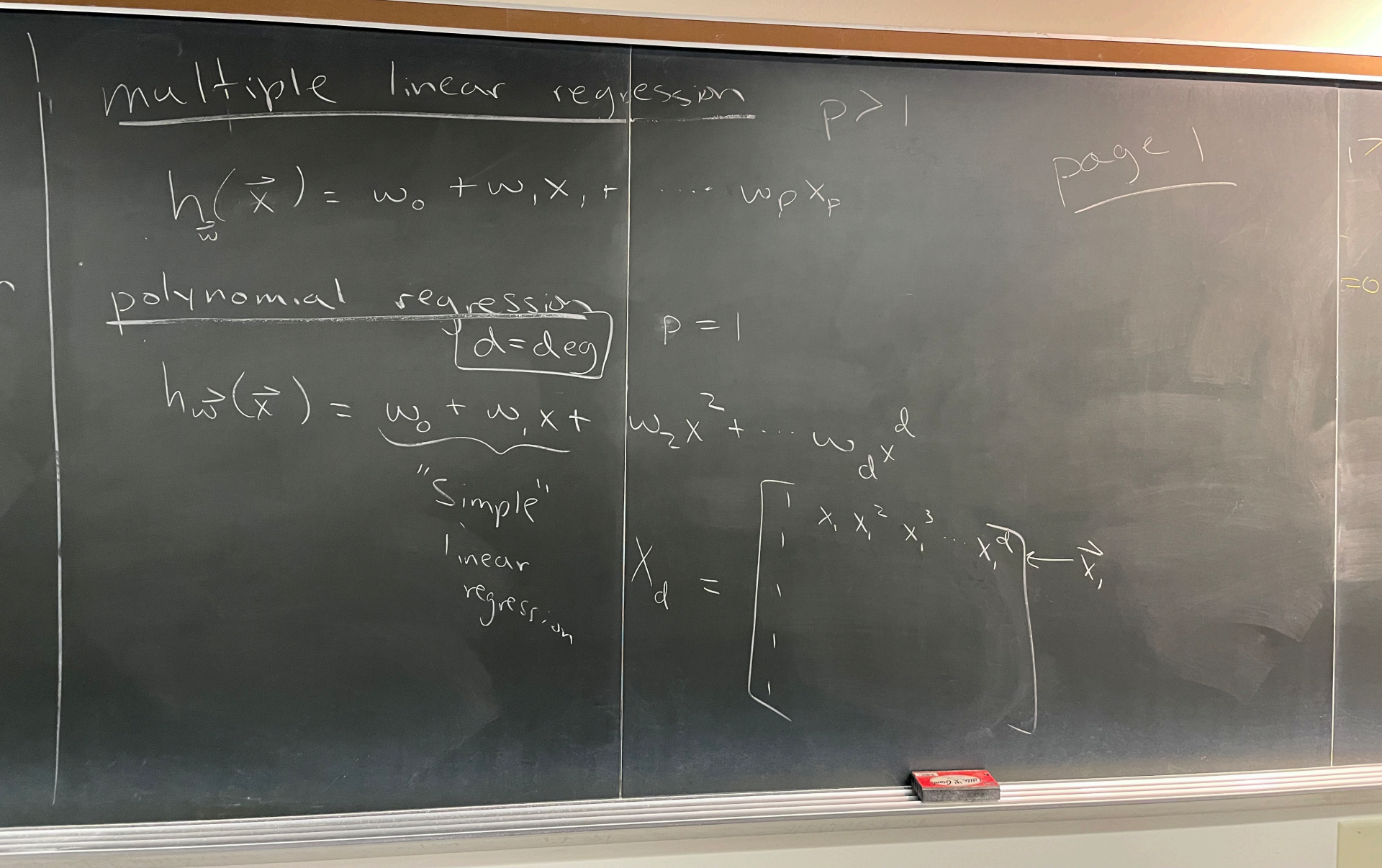


sgd solution to linear regression

Runtime of matrix operations



Multiple linear regression vs. polynomial regression



Handout 10, page 1